ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095961

MULTI-CHANNEL SPEAKER EXTRACTION WITH ADVERSARIAL TRAINING: THE WAVLAB SUBMISSION TO THE CLARITY ICASSP 2023 GRAND CHALLENGE

Samuele Cornell[1,2]*, Zhong-Qiu Wang*[2]*, Yoshiki Masuyama*[3,2]*, Shinji Watanabe*[2]
Manuel Pariente[4]*, Nobutaka Ono*[3]*, Stefano Squartini*[1]

[1]Università Politecnica delle Marche, Italy [2]Carnegie Mellon University, USA
[3]Tokyo Metropolitan University, Japan [4]Pulse Audition, France
{cornellsamuele, wang.zhongqiu41}@gmail.com

ABSTRACT

In this work we detail our submission to the Clarity ICASSP 2023 grand challenge, in which participants have to develop a strong target speech enhancement system for hearing-aid (HA) devices in noisy-reverberant environments. Our system builds on our previous submission at the Second Clarity Enhancement Challenge (CEC2): iNeuBe-X, which consists in an iterative neural/conventional beamforming enhancement pipeline, guided by an enrollment utterance from the target speaker. This model, which won by a large margin the CEC2, is an extension of the state-of-the-art TF-GridNet model for multi-channel, streamable target-speaker speech enhancement. Here, this approach is extended and further improved by leveraging generative adversarial training, which we show proves especially useful when the training data is limited. Using only the official 6k training scenes data, our best model achieves 0.80 hearing-aid speech perception index (HASPI) and 0.41 hearing-aid speech quality index (HASQI) scores on the synthetic evaluation set. However, our model generalized poorly on the semi-real evaluation set. This highlights the fact that our community should focus more on real-world evaluation and less on fully synthetic datasets.

1. INTRODUCTION

The Clarity ICASSP 2023 (CEC2023) grand challenge inherits largely from the previous CEC2 challenge [1]. The dataset is comprised of, respectively, 6k ($\sim$ 10 h) training, 2.5k ($\sim$ 4 h) development multi-channel simulated mixtures. Each mixtures features a listener wearing a 6-microphone (3+3) behind-the-ears HA, a target speaker and up to three different interferers. These interferers could be competing speakers or noise sources. Crucially, the listener rotates their head towards the target, with an unknown random looking direction. As in CEC2, this data is challenging, considering also that models should be causal with a maximum algorithmic latency of 5 ms. In the development set the average mixture scale-invariant signal-to-distortion ratio (SI-SDR) [2] computed against the anechoic target signal is -12.3 dB. A novelty with respect to CEC2 is the fact that a second evaluation set is provided, recorded in a real-world environment, as well as a re-generated synthetic evaluation set. Each of these consists of 1.5k mixtures ($\sim$ 2.4 h), and the semi-real one was recorded in a real room using a 1st order ambisonic microphone. The interferers are simulated using loudspeakers and are synthetically mixed with the clean reverberant target speaker utterance. The other difference compared to CEC2 is that here the listener hearing loss compensation is fixed and ran in the metric evaluation stage. Linear equalization (NAL-R [3]) plus dynamic range compression are used.

2. INEUBE-X SYSTEM OVERVIEW

Our proposed iNeuBe-X system is depicted in Fig. 1 and it is architecturally the same as the one which won the previous CEC2 Challenge [4]. It is based on the iNeuBe (iterative neural/conventional beamforming) [5] framework which won the L3DAS22 ehancement challenge [6]. It employs two multi-microphone input single-microphone output (MISO) DNNs [7] with a conventional beamforming module "sandwiched" in-between. Both DNNs use complex spectral mapping: we feed all channels by simply stacking the input multi-channel STFT real and imaginary (RI) components of all input channels [5]. Then these are trained to predict the complex STFT of the target-speaker anechoic signal S_{target} at the CH1 left of the listener HA array. Thus, target-speaker extraction plus denoising and dereverberation are performed. DNN_1 produces an initial estimate (first iteration $n = 1$) $\hat{S}_1^{(1)}$ which is then fed to a beamforming module. This latter, as in [4], is a causal multichannel wiener filter (MWF) with recursive averaging strategy (forgetting factor 0.5), due to the fact that the listener head rotates. DNN_2 takes in input the original mixture $\boldsymbol{Y}$ multi-channel complex STFT, as well as DNN_1 output and MWF output and produces a refined estimate $\hat{S}_2^{(1)}$. DNN_2 can then be run iteratively [5] (with $\hat{S}_1^{(2)} = \hat{S}_2^{(1)}$), however in our previous work [4], we found it was not worth the additional computation. To help disambiguation between interfering and target speaker, the same DNN_{spk} speaker enrollment module is used as in our previous submission [4]. This module is a small TCNDenseNet [5] with 0.6 million (M) parameters, comprised of only the encoder part, followed by mean-pooling after the temporal convolutional network module. It is used to extract an embedding $X_{\text{adapt}} \in \mathbb{R}^{128}$ from the enrollment utterance. Both DNNs are based on the frame-online speaker-conditioned MISO-TF-GridNet [4] which are an extension of TF-GridNet [8, 9] to multi-channel streamable target enhancement. To condition with X_{adapt} each DNN, feature-wise linear modulation (FiLM) [10] is used at the beginning of each TF-GridNet block.

The whole pipeline uses a 32 kHz sampling rate, 16 ms window, 4 ms stride, and the square-root Hann window. To comply with the latency requirements, the model is trained to predict the current plus future 3 STFT frames as in [11], this leads to a total algorithmic latency of 4 ms. Note that dynamic range compression is applied only in inference and not during training. We use the STFT-domain compressor as in [4] as it led to a small improvement on development set.

2.1. Adversarial Training Fine-Tuning

A key difference with the CEC2 system is the fact that here we propose to fine-tune DNN_2 with a generative adversarial (GAN) training strategy, depicted in Figure 2 and largely borrowed from [12]. The multi-resolution discriminator (MRD) from [12] is employed, which consists of multiple parallel convolutional networks which process the complex STFT of the input signal, each with different window size. Here we use 256, 512, 1024 samples. It outputs the feature maps of each convolutional block plus the final classification logit (a scalar for each window size). A key difference from [12] is that here each network in the MRD is fed X_{adapt}. X_{adapt} is used to condition each network feature maps, prior each convolutional block using independent FiLM layers. The idea is that X_{adapt} will help also the discriminator to disambiguate between the target and interferers, thus improving the "usefulness" of the adversarial loss. Another difference compared to [12] is that here we employ an l_2 loss instead of the hinge loss as the adversarial loss. We found that using l_2 helped considerably and made the training more stable. As in [12] we also use the additional deep feature loss (DFL), it is computed as an l_1 normalized distance between the feature maps (we use all layers as in [12]) when the MRD is fed the enhanced signal vs. when it is fed the target. The gradient

Table 1: Results on development set of CEC2023

Approaches	SI-SDRi (dB)	HASPI	HASQI
mixture CH1 left	0.0	0.24	0.13
oracle CH1 left	∞	0.99	0.73
iNeuBe-X DNN_1	15.34	0.71	0.38
+ DNN_2	17.34	0.78	0.42
+ DNN_2 + GAN	17.24	0.80	0.45
+ DNN_2 + GAN + NAL-R FT	17.89	0.82	0.48

is not propagated back to DNN_{spk}, as this will lead the generator to "cheat" (e.g. by hiding the speaker id) and overcome the MRD.

2.2. System Configuration and Training

For both DNNs we use the same parameters as in [4] and training procedure as well as optimizer configuration, loss functions used etc. DNN_1 is trained first, then DNN_2 is trained, with DNN_1 frozen. We highlight only the differences here due to space limitations. Here we used only the official 6k training scenes in training with no external data for our submission. This is because in CEC2023 the rules prohibited the use of additionally generated data for the main submission. A crucial difference is also the fact that we used the GAN fine-tuning step described previously, prior to the NAL-R fine-tuning step (unchanged from [4]). In the GAN fine-tuning, the adversarial and DFL losses were given a weight of 0.01 and 0.1 respectively, the supervised multi-resolution loss from [4] is still used in this step. For fine-tuning, the model learning rate is $1e-5$, while the MRD's $1e-6$. We perform first GAN fine-tuning and then add also the supervised NAL-R aware fine-tuning loss. Both are performed for 5 epochs each.

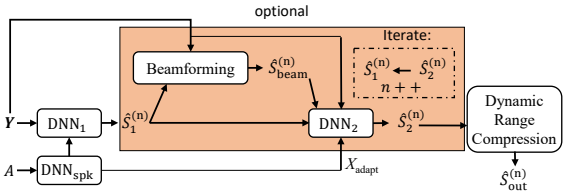

Fig. 1: Overview of the iNeuBe-X framework. Compared to CEC2, here we did not use any compensation module as it is done in the evaluation script.

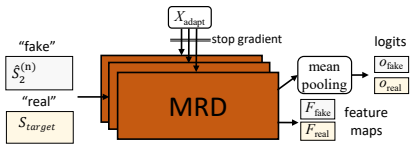

Fig. 2: Overview of the speaker-conditioned multi-resolution discriminator.

3. RESULTS

In Table 1 we report the results obtained on the development set, in terms of SI-SDR, hearing-aid speech perception and quality indexes (HASPI) [13] and (HASQI) [14] respectively. These two latter are computed taking into account the fixed compensation + dynamic compression provided by the organizers. We can see that the upper bound (oracle CH1 left) for HASQI (anechoic target at HA array CH1 left) using the pre-provided compensation is rather low. For comparison, we also report the results with no enhancement at all (mixture CH1 left). The largest gain, is observed when DNN_2 is added. The proposed GAN adversarial training is especially effective towards HASQI, even if degrades a bit SI-SDR. Adding the NAL-R fine-tuning strategy [4] further improves the results for all metrics. In general, the results here are much worse than ones in our previous work [4] due to the much smaller training set and the fact that we did not use our compensation (across which we can back-propagate in the NAL-R fine-tuning step). In particular, regarding HASQI, better compensation strategies needs to be devised in order to improve it, as the oracle results demonstrates.

Table 2: Results on evaluation set of CEC2023.

Approaches	Synth		Real	
	HASPI	HASQI	HASPI	HASQI
mixture CH1 Left	0.26	0.13	0.18	0.12
ours (submitted)	0.80	0.41	0.29	0.11
ours + additional	0.85	0.46	0.33	0.11

In Table 2 we report the results, obtained on the evaluation set, courtesy of CEC2023 organizers. We can see that our proposed method performs reasonably well on the synthetic evaluation. However it fails to reach satisfactory performance on the semi-real evaluation (HASQI degrades with respect to no enhancement). We ran an additional experiment (ours + additional) after CEC2023 end. In particular, we added to the training material 2k additional scenes simulated from the same data but with 1st ambisonic order (instead of 6-th), to match the one of the real-world recordings. We can see that the performance increases in both evaluation sets, but more on the synthetic than the real-world one, on which it remains poor. This suggest that the ambisonic order is not the main source of mismatch.

4. CONCLUSIONS

In this short paper we presented our submission to the Clarity 2023 ICASSP Grand Challenge. It builds considerably upon our previous submission [4], but here we devised an additional adversarial training strategy that seems helpful especially regarding the HASQI score. This model ranked in the top five with scores promising scores on the synthetic evaluation, but the model fails to enhance on the real-world evaluation data. Further work is needed to assess the causes and potential solutions to this generalization problem.

5. REFERENCES

[1] S. Graetzer and et al., "Clarity: machine learning challenges to revolutionise hearing device processing," in *Proceedings of Forum Acusticum 2020*, 2020, pp. 3495–3497.

[2] J. Le Roux and et al., "Sdr–half-baked or well done?" in *Proc. ICASSP*, 2019.

[3] D. Byrne and H. Dillon, "The National Acoustic Laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and hearing*, vol. 7, no. 4, pp. 257–265, 1986.

[4] S. Cornell and et al., "Multi-channel target speaker extraction with refinement: The wavlab submission to the second clarity enhancement challenge," in *The Second Clarity Enhancement Challenge Workshop*, 2022. [Online]. Available: https://arxiv.org/abs/2302.07928

[5] Y.-J. Lu and et al., "Towards low-distortion multi-channel speech enhancement: The ESPNET-SE submission to the L3DAS22 challenge," in *Proc. ICASSP*, 2022, pp. 9201–9205.

[6] E. Guizzo and et al., "L3DAS22 challenge: Learning 3D audio sources in a real office environment," in *Proc. ICASSP*, 2022.

[7] Z.-Q. Wang and et al., "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2001–2014, 2021.

[8] ——, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *accepted at ICASSP 2023*, 2022.

[9] ——, "Tf-gridnet: Integrating full-and sub-band modeling for speech separation," *submitted to IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2022.

[10] E. Perez and et al., "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, vol. 32, no. 1, 2018.

[11] Z.-Q. Wang and et al., "STFT-domain neural speech enhancement with very low algorithmic latency," *arXiv preprint arXiv:2204.09911*, 2022.

[12] A. Défossez and et al., "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[13] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (hasqi) version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[14] ——, "The hearing-aid speech perception index (haspi) version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.